HGG Advances

ARTICLE

Pleiotropy-guided transcriptome imputation from normal and tumor tissues identifies candidate susceptibility genes for breast and ovarian cancer

Siddhartha P. Kar,[1,2,65,*] Daniel P.C. Considine,[3,65] Jonathan P. Tyrer,[4] Jasmine T. Plummer,[5,6] Stephanie Chen,[5,6] Felipe S. Dezem,[5,6] Alvaro N. Barbeira,[7] Padma S. Rajagopal,[8] Will T. Rosenow,[9] Fernando Moreno,[10] Clara Bodelon,[11] Jenny Chang-Claude,[12,13] Georgia Chenevix-Trench,[14] Anna deFazio,[15,16] Thilo Dörk,[17] Arif B. Ekici,[18,19] Ailith Ewing,[20,21] George Fountzilas,[22] Ellen L. Goode,[23] Mikael Hartman,[24,25] Florian Heitz,[26,27] Peter Hillemanns,[28] Estrid Høgdall,[29,30] Claus K. Høgdall,[31] Tomasz Huzarski,[32,33] Allan Jensen,[34] Beth Y. Karlan,[35] Elza Khusnutdinova,[36,37] Lambertus A. Kiemeney,[38] Susanne K. Kjaer,[29,39] Rüdiger Klapdor,[28] Martin Köbel,[40] Jingmei Li,[24,41] Clemens Liebrich,[42] Taymaa May,[43] Håkan Olsson,[44] Jennifer B. Permuth,[45] Paolo Peterlongo,[46] Paolo Radice,[47] Susan J. Ramus,[48,49] Marjorie J. Riggan,[50] Harvey A. Risch,[51] Emmanouil Saloustros,[52]

(Author list continued on next page)

Summary

Familial, sequencing, and genome-wide association studies (GWASs) and genetic correlation analyses have progressively unraveled the shared or pleiotropic germline genetics of breast and ovarian cancer. In this study, we aimed to leverage this shared germline genetics to improve the power of transcriptome-wide association studies (TWASs) to identify candidate breast cancer and ovarian cancer susceptibility genes. We built gene expression prediction models using the PrediXcan method in 681 breast and 295 ovarian tumors from The Cancer Genome Atlas and 211 breast and 99 ovarian normal tissue samples from the Genotype-Tissue Expression project and integrated these with GWAS meta-analysis data from the Breast Cancer Association Consortium (122,977 cases/105,974 controls) and the Ovarian Cancer Association Consortium (22,406 cases/40,941 controls). The integration was achieved through application of a pleiotropy-guided conditional/conjunction false discovery rate (FDR) approach in the setting of a TWASs. This identified 14 candidate breast cancer susceptibility genes spanning 11 genomic regions and 8 candidate ovarian cancer susceptibility genes spanning 5 genomic regions at conjunction FDR < 0.05 that were >1 Mb away from known breast and/or ovarian cancer susceptibility loci. We also identified 38 candidate breast cancer susceptibility genes and 17 candidate ovarian cancer susceptibility genes at conjunction FDR < 0.05 at known breast and/or ovarian susceptibility loci. The 22 genes identified by our cross-cancer analysis represent promising candidates that further elucidate the role of the transcriptome in mediating germline breast and ovarian cancer risk.

Introduction

The last three decades have witnessed major advances in our understanding of the shared inherited genetic basis of breast and ovarian cancer. The identification of rare inherited mutations in *BRCA1* (MIM: 113705)[1] and *BRCA2* (MIM: 600185)[2] that confer high risks of developing both breast and ovarian cancer has directly opened up the identification of oncogenic mechanisms leading to the development of poly ADP ribose polymerase

[1]Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol, UK; [2]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK; [3]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK; [4]Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK; [5]Center for Bioinformatics and Functional Genomics, Department of Biomedical Science, Cedars-Sinai Medical Center, Los Angeles, CA, USA; [6]Department of Biomedical Sciences, Cedars Sinai Medical Center, Los Angeles, CA, USA; [7]Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, IL, USA; [8]Section of Hematology/Oncology, Department of Medicine, University of Chicago, Chicago, IL, USA; [9]Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA; [10]Department of Oncology, Hospital Clínico San Carlos, Madrid, Spain; [11]Divison of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA; [12]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany; [13]Cancer Epidemiology Group, University Cancer Center Hamburg (UCCH), University Medical Center Hamburg-Eppendorf, Hamburg, Germany; [14]Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, QLD, Australia; [15]Centre for Cancer Research, The Westmead Institute for Medical Research, The University of Sydney, Sydney, NSW, Australia; [16]Department of Gynaecological Oncology, Westmead Hospital, Sydney, NSW, Australia; [17]Gynaecology Research Unit, Hannover Medical School, Hannover, Germany; [18]Institute of Human Genetics, University Hospital Erlangen, Erlangen, Germany; [19]Friedrich-Alexander University Erlangen-Nuremberg, Comprehensive Cancer Center Erlangen, Erlangen, Germany; [20]MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK; [21]Cancer Research UK Edinburgh Centre, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK; [22]Laboratory of Molecular Oncology, Hellenic Foundation for Cancer Research, Aristotle University of Thessaloniki School of Medicine, Thessaloniki, Greece; [23]Department of Quantitative Health Sciences, Division of Epidemiology, Mayo Clinic, Rochester,

(Affiliations continued on next page)

Jacques Simard,[53] Lukasz M. Szafron,[54] Linda Titus,[55] Cheryl L. Thompson,[56] Robert A. Vierkant,[57] Stacey J. Winham,[58] Wei Zheng,[59] Jennifer A. Doherty,[60] Andrew Berchuck,[61] Kate Lawrenson,[5,62] Hae Kyung Im,[7] Ani W. Manichaikul,[9,63] Paul D.P. Pharoah,[3,4] Simon A. Gayther,[5] and Joellen M. Schildkraut[64]

inhibitor therapy.[3] The findings from genome-wide association studies (GWASs) have demonstrated that there is a strong genetic correlation between breast and ovarian cancer[4] and have identified several genomic regions containing common (minor allele frequency > 1%) variants that confer risk of developing both breast and ovarian cancer.[5,6]

Transcriptome-wide association studies (TWASs) represent the latest study design for the identification of disease-associated susceptibility genes. TWASs involve establishing robust multi-variant models for the component of somatic (normal or tumor) gene expression that is regulated by germline genetic variation in a smaller dataset where both germline genotype and somatic transcriptomic data are available. These models are then used to impute the germline genetically regulated component of gene expression into a larger GWAS dataset where measured gene expression is unavailable but that offers significantly improved power to identify genes associated with disease risk where such risk may be mediated by expression. Moving from single variants (GWASs) to genes (TWASs) as the unit of association reduces the multiple testing burden. The use of gene expression provides a readily accessible readout of the functional basis of the identified association in contrast to GWAS-identified risk variants that predominantly reside in non-coding regions of the genome.[7]

PrediXcan is a method developed recently for conducting TWASs.[8] TWAS methods have been applied to single cancer types before, including breast cancer[9,10] and ovarian cancer.[11,12] Here we present an application of PrediXcan, and indeed broadly of TWASs, in the pleiotropic cross-cancer setting. We used the normal and tumor breast- and ovary-specific gene expression and matched germline genotype datasets to generate tissue-specific PrediXcan models and first imputed these models into GWAS data for the corresponding cancers (i.e., from breast-tissue-derived models into breast cancer GWASs and likewise for the ovarian models). We then imputed models across cancer types (i.e., from breast-tissue-derived models into ovarian cancer GWASs and vice versa). Finally, we implemented a powerful conjunction false discovery rate (FDR) approach[13,14] that has been applied previously to GWASs,[15–18] but not to TWASs, to leverage the combined GWAS sample of over 145,000 breast and ovarian cancer cases. We identify candidate breast and ovarian cancer susceptibility genes in regions not previously implicated by GWAS or TWAS analyses of these cancers.

MN, USA; [24]Department of Surgery, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore, Singapore; [25]Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, Singapore, Singapore; [26]Department of Gynecology and Gynecologic Oncology, Kliniken Essen-Mitte/Evang., Essen, Germany; [27]Department of Gynecology, Center for Oncologic Surgery, Charité Campus Virchow-Klinikum, Berlin, Germany; [28]Department of Gynecology and Obstetrics, Hannover Medical School, Hannover, Germany; [29]Department of Virus, Lifestyle, and Genes, Danish Cancer Society Research Center, Copenhagen, Denmark; [30]Molecular Unit, Department of Pathology, Herlev Hospital, University of Copenhagen, Copenhagen, Denmark; [31]The Juliane Marie Centre, Department of Gynecology, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark; [32]Department of Genetics and Pathology, International Hereditary Cancer Center, Pomeranian Medical University, Szczecin, Poland; [33]Department of Genetics and Pathology, University of Zielona Góra, Zielona Góra, Poland; [34]Department of Lifestyle, Reproduction, and Cancer, Danish Cancer Society Research Center, Copenhagen, Denmark; [35]David Geffen School of Medicine, Department of Obstetrics and Gynecology, University of California at Los Angeles, Los Angeles, CA, USA; [36]Institute of Biochemistry and Genetics, Ufa Federal Research Centre of the Russian Academy of Sciences, Ufa, Russia; [37]Department of Genetics and Fundamental Medicine, Bashkir State University, Ufa, Russia; [38]Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, the Netherlands; [39]Department of Gynaecology, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark; [40]Department of Pathology and Laboratory Medicine, University of Calgary, Foothills Medical Center, Calgary, AB, Canada; [41]Genome Institute of Singapore, Human Genetics, Singapore, Singapore; [42]Department of Obstetrics and Gynecology, Klinikum Wolfsburg, Wolfsburg, Germany; [43]Division of Gynecologic Oncology, University Health Network, Princess Margaret Hospital, Toronto, ON, Canada; [44]Division of Oncology, Department of Clinical Sciences, Lund University, Lund, Sweden; [45]Departments of Cancer Epidemiology and Gastrointestinal Oncology, Moffitt Cancer Center and Research Institute, Tampa, FL, USA; [46]Genome Diagnostics Program, IFOM-The FIRC Institute of Molecular Oncology, Milan, Italy; [47]Unit of Molecular Bases of Genetic Risk and Genetic Testing, Department of Research, Fondazione IRCCS Istituto Nazionale dei Tumori (INT), Milan, Italy; [48]School of Women's and Children's Health, Faculty of Medicine, University of New South Wales, Sydney, NSW, Australia; [49]Adult Cancer Program, Lowy Cancer Research Centre, University of New South Wales, Sydney, NSW, Australia; [50]Department of Obstetrics and Gynecology, Duke University Medical Center, Durham, NC, USA; [51]Department of Chronic Disease Epidemiology, Yale School of Public Health, New Haven, CT, USA; [52]Department of Oncology, University Hospital of Larissa, Larissa, Greece; [53]Genomics Center, Centre Hospitalier Universitaire de Québec - Université Laval Research Center, Québec City, QC, Canada; [54]Maria Sklodowska-Curie National Research Institute of Oncology, Warsaw, Poland; [55]Muskie School of Public Service, University of Southern Maine, Portland, ME, USA; [56]Department of Nutrition, Case Western Reserve University, Cleveland, OH, USA; [57]Department of Quantitative Health Sciences, Division of Clinical Trials and Biostatistics, Mayo Clinic, Rochester, MN, USA; [58]Department of Quantitative Health Sciences, Division of Computational Biology, Mayo Clinic, Rochester, MN, USA; [59]Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN, USA; [60]Huntsman Cancer Institute, Department of Population Health Sciences, University of Utah, Salt Lake City, UT, USA; [61]Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, Duke University Medical Center, Durham, NC, USA; [62]Women's Cancer Program at the Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA; [63]Department of Public Health Sciences, University of Virginia, Charlottesville, VA, USA; [64]Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA, USA
[65]These authors contributed equally
*Correspondence: siddhartha.kar@bristol.ac.uk
https://doi.org/10.1016/j.xhgg.2021.100042.

Material and methods

Matched germline genotype: normal/tumor gene expression datasets

We used data for 211 normal breast tissue samples and 99 normal ovarian tissue samples from the Genotype-Tissue Expression (GTEx) project (version 7 release).[19] Germline genotypes in the GTEx data had been called from whole-genome sequencing (Illumina HiSeq X), and gene expression was profiled using RNA-sequencing (Illumina TruSeq). We also used data from 681 breast cancer[20] and 295 high-grade serous ovarian cancer (HGSOC)[21] cases from The Cancer Genome Atlas (TCGA) network. Germline genotypes in the TCGA data had been called from genotyping arrays (Affymetrix SNP 6.0), and gene expression was profiled using RNA-sequencing (Illumina HiSeq 2000). Imputation of TCGA germline genotypes using the 1000 Genomes version 5 reference panel was performed as described previously.[22,23] TCGA sample sizes reported here refer to only those samples that had >95% European ancestry. Ancestry was estimated using the Local Ancestry in adMixed Populations tool (LAMP version 2.5).[24] Downstream PrediXcan modeling (described below) used variants imputed with quality > 0.8 that had a minor allele frequency $> 5\%$ in TCGA datasets.

Genome-wide association datasets

Summary statistics from genome-wide association meta-analyses were obtained from the Breast Cancer Association Consortium (BCAC)[22] and the Ovarian Cancer Association Consortium (OCAC).[23] The breast cancer susceptibility data were based on 122,977 cases and 105,974 controls, including 21,468 estrogen receptor (ER)-negative cases. The ovarian cancer susceptibility data were based on 22,406 epithelial ovarian cancer cases and 40,941 controls, including 13,037 HGSOC cases. We harmonized the signs of the effect size estimates and aligned them to the same effect allele in the breast and ovarian cancer GWAS datasets. We retained 9,530,997 variants with minor allele frequency $> 1\%$ and imputation quality > 0.4 in both datasets for S-PrediXcan analyses. All individuals in these studies were of genetically inferred European ancestry.

PrediXcan model development and S-PrediXcan analyses

We built genetically regulated gene expression prediction models using the elastic net regularization approach implemented in PrediXcan and validated these models using tenfold cross-validation.[8] Essentially, this generates a list of variants for each gene where model construction is successful and each variant in the list is assigned a weight reflecting its influence on its target gene expression. Genes with models where the nested tenfold cross-validated correlation between predicted and actual levels of expression was >10% (predictive performance $r^2 > 0.01$) and p value of the correlation test was <0.05 were retained. These models were adjusted for the latent determinants of gene expression variation (referred to hereafter as PEER factors), which were identified using the Probabilistic Estimation of Expression Residuals (PEER; version 1.3) method.[25] We adjusted for 60 and 45 PEER factors for TCGA breast and ovarian cancer data, respectively. The choice of these numbers is a function of sample size and consistent with recommendations.[8,25] *ESR1* expression was also included as a covariate in the construction of breast cancer models to account for ER status and its influence on the expression of individual genes. For the GTEx version 7 datasets, we downloaded pre-computed PrediXcan models from predictdb.org. Our pipeline for processing the TCGA datasets, including the application of PEER factors, was designed to be consistent with the pipeline used to generate the pre-computed GTEx PrediXcan models. S-PrediXcan refers to the application of the PrediXcan gene expression models, specifically the variant weights from elastic net combined into multi-variant gene-level instruments, to summary statistics GWAS datasets and has been described in detail before.[8] The variance of a gene's expression that was explained by the SNPs in its model was calculated as $W' \times G \times W$ (where W is the vector of SNP weights in a gene's model, W' is its transpose, and G is the covariance matrix).

Conditional and conjunction FDR analyses

We obtained p values for association of predicted expression of each gene with breast cancer risk and with ovarian cancer risk. We then computed the FDR for gene-breast cancer risk association conditional on gene-ovarian cancer risk association (as conditional $FDR_{\text{Breast Cancer|Ovarian Cancer}}$). This is the probability that a gene is not associated with breast cancer risk given the p values for association with both breast cancer risk and ovarian cancer risk. The analogous conditional FDR for gene-ovarian cancer risk association was also calculated ($FDR_{\text{Ovarian Cancer|Breast Cancer}}$). Finally, the conjunctional FDR estimate, which is conservatively defined as the maximum of the two conditional FDR values, was computed. This process minimizes the effect of a single phenotype (in this case, breast or ovarian cancer) driving the shared association signal. It allows the power of pleiotropic associations to be tapped for genetic discovery, unlike a traditional FDR approach that is informed solely by the distribution of p values for a single phenotype. We used the R implementation of the conditional FDR method. The conditional and conjunctional FDR method has been described extensively elsewhere[13–18] but not applied before to the TWAS setting. The overall study design is summarized in Figure 1.

Fine-mapped candidate causal risk variant datasets

We examined the overlap between variants in the breast gene expression prediction models and a published list of fine-mapped candidate causal risk variants for breast cancer.[26] This was done to follow up genes that we identified in genomic regions that are known to be associated with breast cancer risk under the intuition that gene-level association signals identified by S-PrediXcan that demonstrate such overlap with fine-mapped variants are likely being driven by the GWAS association signal in the same region.

Fine-mapped candidate causal risk variants lists for breast cancer were obtained from Fachal et al.[26] Briefly, Fachal et al. fine-mapped 150 known breast cancer susceptibility regions using dense genotype data on women participating in the BCAC and in the Consortium of Investigators of Modifiers of BRCA1/2 (CIMBA). Stepwise multinomial logistic regression was used to identify independent association signals in each region. Credible causal variants within each signal were defined as being within a 100-fold likelihood of the top conditional variant to delineate the variants driving the GWAS associations in each region.

We adopted a similar analytic strategy for the ovarian cancer dataset from OCAC. Each genomic region with a genome-wide significant ($p < 5 \times 10^{-8}$) variant was explored to identify additional independent association signals. All variants within a given genomic region were jointly analyzed to evaluate the simultaneous

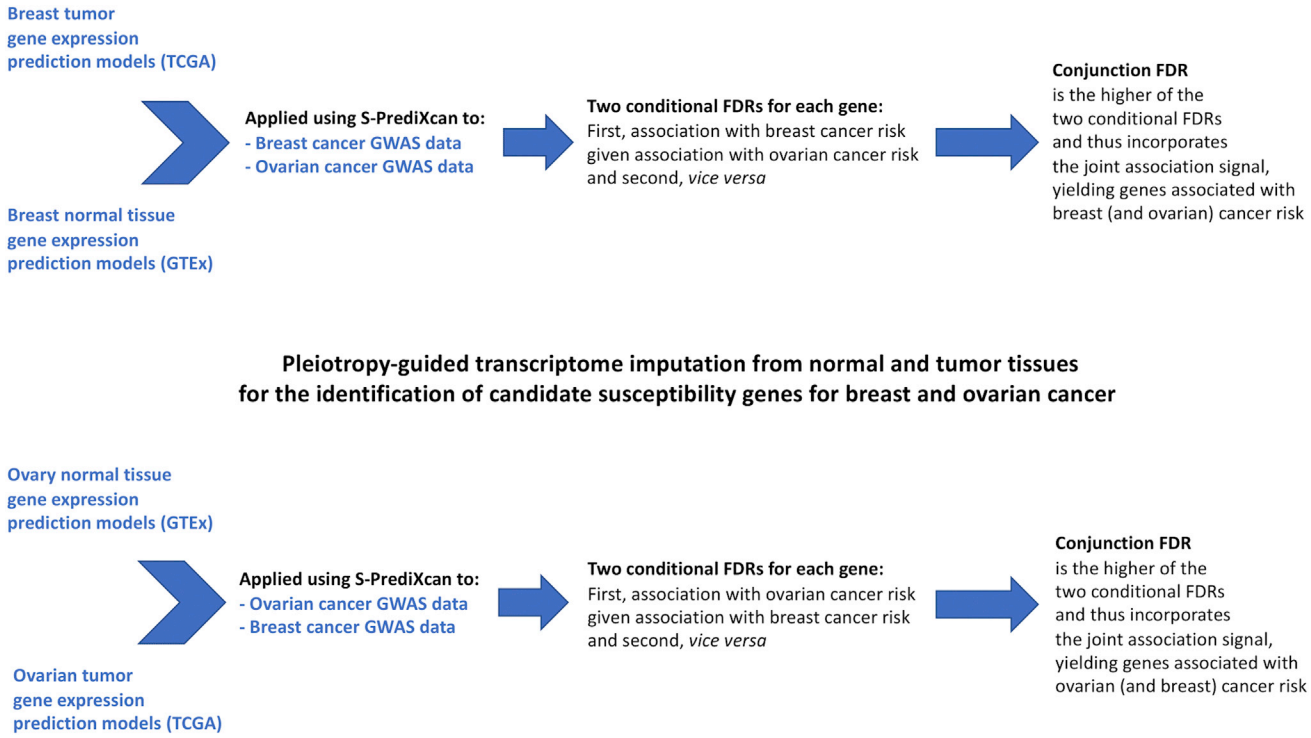

Figure 1. Overview of datasets and analyses in this study
Flowchart providing an overview of the datasets used and the various steps in the analysis. GTEx, Genotype-Tissue Expression project; TCGA, The Cancer Genome Atlas; GWAS, genome-wide association study; FDR, false discovery rate.

effects of multiple variants, using a 1 Mb window centered on the most significant variant, in stepwise conditional models. Given the presence of a genome-wide significant variant in the region, the prior probability of an additional risk variant in the same region is higher than in a region without a genome-wide significant lead variant; therefore, we used a threshold of $p < 1 \times 10^{-5}$ to identify additional independent association signals. All variants in each region were ranked by the likelihood of association with ovarian cancer based on p values. The likelihood of each variant was then compared with the likelihood of the lead variant in the region based on the primary association analysis for primary signals and the conditional association analysis for conditional signals. Variants with odds > 1:100 compared with the lead variant (corresponding to a p value 100 times larger than the most significant p value[27]) were selected as credible causal variants.

Results

Development of tissue/tumor-specific gene expression prediction models

We built genetically regulated gene expression predictor models using matched germline genotype and tumor gene expression data from TCGA by applying elastic net regularization as implemented in the PrediXcan software. Genes with models where the nested tenfold cross-validated correlation between predicted and actual levels of expression was >10% (predictive performance $r^2 > 0.01$) and p value of the correlation test was <0.05 were retained in line with best practice quality control recommendations by the developers of PrediXcan.[8] We constructed and evaluated predictor models that met these criteria for 4,457 genes based on 681 TCGA breast tumor samples and for 2,705 genes based on 295 TCGA ovarian tumor samples. We obtained pre-computed genetically regulated gene expression predictor models that met the same criteria (predictive performance $r^2 > 0.01$; correlation test $p < 0.05$) in matched germline genotype and normal tissue gene expression data from the GTEx Project. Specifically, the pre-computed data included 5,274 genes modeled based on 211 GTEx breast tissue samples and 3,034 genes modeled based on 99 GTEx ovarian tissue samples. The variance of a gene's expression explained by SNPs in its model was, on average, lower in tumors and higher in normal tissues (mean [standard deviation] for TCGA breast cancer: 0.04 [0.07]; TCGA ovarian cancer: 0.05 [0.06]); GTEx breast: 0.09 [0.09]; and GTEx ovary: 0.15 [0.13]), likely reflecting the relatively smaller influence of germline genetic variation on tumor gene expression compared to its impact on normal tissue gene expression. Prediction performance as measured by the cross-validated correlation of the tissue model's correlation to the gene's measured transcriptome was, in general, substantially better for the normal tissue models than the tumor tissue models (Figure S1).

Imputation of gene expression into GWAS and pleiotropy-guided FDR control

We used the GTEx normal breast-tissue-derived prediction models to impute genetically regulated gene expression in

a genome-wide association meta-analysis involving 122,977 breast cancer cases and 105,974 controls using S-PrediXcan. We tested for association between imputed gene expression and breast cancer risk. We also used the same GTEx breast-tissue-based models to impute gene expression in a genome-wide association meta-analysis of 22,406 ovarian cancer cases and 40,941 controls and test for association between imputed expression and ovarian cancer risk. For these two steps, we applied the conditional FDR method to the S-PrediXcan gene-level association p values to correct for testing 5,274 genes in each analysis. This yielded two conditional FDR values: one for association with breast cancer risk given association with ovarian cancer risk and the other for association with ovarian cancer risk given association with breast cancer risk. Finally, we took the larger of the two values for each gene as a conservative estimate of its conjunction FDR to identify candidate breast cancer susceptibility genes at conjunction FDR < 0.05. We refer to these genes as candidate breast cancer susceptibility genes because they were identified on the basis of gene expression predictor models derived from breast tissue. However, the conditional-conjunction FDR analysis effectively borrowed information from pleiotropic associations with inherited susceptibility to a second cancer type (in this case ovarian cancer) in addition to the primary cancer type (breast cancer), and these genes may be considered as risk genes for the second cancer as well. These steps were repeated for three other ordered combinations of datasets: TCGA breast tumor tissue-breast cancer GWAS-ovarian cancer GWAS to identify candidate breast cancer susceptibility genes; GTEx normal ovarian tissue-ovarian cancer GWAS-breast cancer GWAS and TCGA ovarian tumor tissue-ovarian cancer GWAS-breast cancer GWAS to identify candidate ovarian cancer susceptibility genes. We also replaced the overall breast cancer GWASs and all invasive ovarian cancer GWASs used in the four dataset combinations described above with ER-negative breast cancer GWASs (21,468 cases/105,974 controls) and HGSOC GWASs (13,037 cases/22,406 controls), respectively. This helped identify additional candidate breast and ovarian cancer susceptibility genes driven by subtype-specific associations at conjunction FDR < 0.05.

For each gene, coverage was defined as the percentage of the number of variants included in its expression prediction model that were also captured in the genome-wide association meta-analysis. The coverage was $\geq$80% for at least 93% of the genes in each of the four matched germline genotype and normal or tumor gene expression datasets used to build the predictor models, indicating that for most genes, most of the corresponding model variants available were used. In each ordered analytic combination of datasets (e.g., GTEx normal breast tissue-breast cancer GWAS-ovarian cancer GWAS) we observed that, in general, for progressively smaller S-PrediXcan p values of the second cancer type, the true discovery rate for association with the primary cancer type approached 100% at progressively larger S-PrediXcan p values for the primary cancer type (Figure 2; Figure S2). This was consistent with substantial shared gene-level associations for breast and ovarian cancer risk and these shared signals being tapped by the conditional-conjunction FDR method to power candidate susceptibility gene discovery.

Identification of candidate breast cancer and ovarian cancer susceptibility genes

We identified 14 candidate breast cancer susceptibility genes at the conjunction FDR < 0.05 threshold (Table 1; Table S1). The 14 genes were distributed between 11 genomic regions >1 Mb apart from each other (Table 1). These genes have not been reported as susceptibility genes in any prior TWAS of breast cancer risk and are >1 Mb away from published genome-wide significant lead variants for breast cancer susceptibility.[28] For ovarian cancer, we identified 8 candidate susceptibility genes at conjunction FDR < 0.05 (Table 2; Table S2). The 8 genes were located across 5 genomic regions >1 Mb apart from each other (Table 2). These genes have not been reported as candidate risk genes in any previously reported TWASs of ovarian cancer risk and are >1 Mb away from published genome-wide significant lead variants for ovarian cancer susceptibility.[23]

Candidate breast cancer and ovarian cancer susceptibility genes at known GWAS loci

We identified 38 candidate breast cancer susceptibility genes that were located within 1 Mb of a published lead variant associated at genome-wide significance with breast cancer risk (Table S3).[28] Four of the 38 genes have also been reported in previously published TWASs (Table S3).[9,10] The 38 genes were spread across 12 genomic regions >1 Mb apart from each other. Overlaying fine-mapped candidate causal breast cancer risk variants on breast gene expression predictor model variants showed that for 21/38 (55%) genes, the prediction model variants included at least one fine-mapped candidate causal variant (Tables S3 and S4). This suggested that, for these genes, the GWAS association signal was driving the S-PrediXcan signal. We also identified three additional genes that were >1 Mb away from known GWAS loci that have previously been reported as TWAS loci for breast cancer risk (Table S3).[9,10]

For ovarian cancer, we identified 17 candidate susceptibility genes that were located within 1 Mb of a published lead variant associated at genome-wide significance with ovarian cancer risk (Table S5).[23] Six of these genes have also been reported in a previously published TWAS for ovarian cancer (Table S5).[11,12] The 17 genes span 5 different genomic regions >1 Mb apart. Overlaying fine-mapped candidate causal ovarian cancer risk variants onto the ovarian gene expression predictor model variants showed that for 12/17 (71%) genes, the prediction model variants included at least one fine-mapped candidate causal variant (Tables S5 and S6), suggesting that for these genes the GWAS association signal underpinned the S-PrediXcan signal.

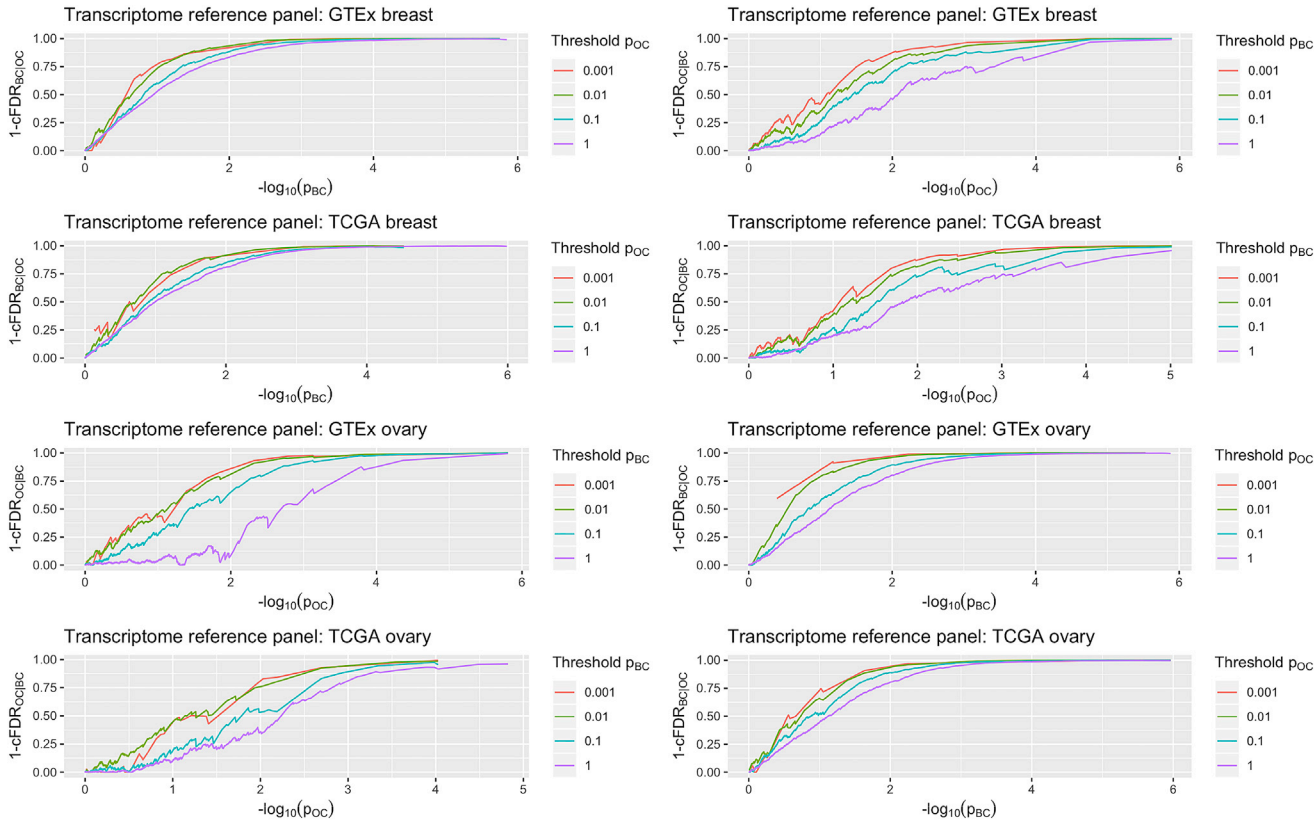

Figure 2. True discovery rate of S-PrediXcan associations for each cancer stratified by associations with the other cancer
True discovery rate against the negative logarithm (base 10) of the p value for each cancer for subsets of genes based on strength of association with the other cancer. The y axis of each plot is the true discovery rate, which is defined as 1 − conditional FDR (cFDR). For a given ordered analytic combination of datasets (e.g., GTEx normal breast tissue as transcriptome reference panel-breast cancer GWAS-ovarian cancer GWAS, plotted in the upper left corner) we observed that, in general, for progressively smaller S-PrediXcan p values of the second cancer type (indicated by the key "Threshold p" next to each plot), the true discovery rate (y axis) for association with the primary cancer type approached 100% at progressively larger S-PrediXcan p values for the primary cancer type (x axis; negative logarithm [base 10] of the p values). Only p values $> 10^{-6}$ are plotted on the x axis. BC, overall breast cancer risk; OC, all invasive ovarian cancer risk.

Discussion

In this study, we used the conditional and conjunctional FDR as a tool to systematically improve the power of breast cancer and ovarian cancer candidate susceptibility gene discovery in a PrediXcan-based TWAS. While gene expression prediction models based on multiple tissue types have been the more common approach to improving TWAS power,[11,29] the conditional/conjunction FDR approach gains power through the incorporation of multiple related GWAS datasets into a TWAS analysis. We investigated the shared inherited genetic basis of these two cancer types by integrating normal and tumor-tissue-specific transcriptomic datasets with large-scale genome-wide association meta-analysis findings for susceptibility to breast cancer and ovarian cancer. We identified 11 genomic regions associated with breast cancer risk and five regions linked to ovarian cancer risk.

We identified 14 candidate breast cancer susceptibility genes (Table 1). Many of these genes have a strong biological rationale for involvement in breast carcinogenesis and are in or near genomic regions associated with other cancer types or potential cancer risk factors. For example, the *ZNF276* (MIM: 608460) intronic variant rs12925026 is associated at genome-wide significance with non-melanoma skin cancer.[30] *ZNF276* overlaps *FANCA* (MIM: 607139) in a tail-to-tail manner.[31] The genetically regulated predictor model for *ZNF276* expression was fit using gene expression measured in GTEx breast tissues, but neither this dataset nor any of the other datasets could capture a predictor model for *FANCA* expression. *FANCA* encodes one of eight subunits that together form the core Fanconi Anemia (FA) complex that repairs blockages in DNA replication due to cross-linking.[32] Several members of the FA family of proteins have been implicated in breast and ovarian cancer predisposition, including *BRCA1* (*FANCS*), *BRCA2* (*FANCD1*), *BRIP1* (MIM: 605882) (*FANCJ*), *PALB2* (MIM: 610355) (*FANCN*), *RAD51C* (MIM: 602774) (*FANCO*), and *FANCM* (MIM: 609644), and it is possible that *FANCA* may represent another or possibly the true target breast cancer susceptibility gene in this region, given this biological function and its overlap with *ZNF276*.[32,33] *ZNF276* in its own right has also been implicated as a candidate tumor suppressor gene in breast cancer,[31] and

Table 1. Candidate breast cancer susceptibility genes identified by pleiotropy-guided S-PrediXcan analysis

Gene	Genomic region	p value BC	p value OC	Conditional FDR BC|OC	Conditional FDR OC|BC	Conjunction FDR
Transcriptome reference panel: GTEx breast (normal)|primary GWAS: overall BC risk (second GWAS: all invasive OC risk)						
ZSCAN29	15q15.3	1.8E−04	9.1E−04	4.1E−04	6.0E−03	6.0E−03
STRCP1	15q15.3	1.6E−03	1.8E−03	4.1E−03	1.9E−02	1.9E−02
AC011330.5	15q15.3	5.4E−04	2.5E−03	1.8E−03	2.2E−02	2.2E−02
STRC	15q15.3	1.4E−04	3.8E−03	6.1E−04	2.2E−02	2.2E−02
ZNF276	16q24.3	3.7E−06	4.7E−03	2.4E−05	2.2E−02	2.2E−02
RGS19	20q13.33	1.1E−03	5.4E−03	4.3E−03	4.3E−02	4.3E−02
RNFT1	17q23.1	2.4E−04	8.7E−03	1.3E−03	4.7E−02	4.7E−02
C15orf65	15q21.3	2.2E−03	5.9E−03	8.2E−03	4.7E−02	4.7E−02
Transcriptome reference panel: TCGA breast (tumor)|primary GWAS: overall BC risk (second GWAS: all invasive OC risk)						
GMNC	3q28	2.6E−03	1.2E−03	6.1E−03	2.0E−02	2.0E−02
ESRP2	16q22.1	1.9E−02	9.6E−04	4.3E−02	3.3E−02	4.3E−02
BHLHA15	7q21.3	8.5E−05	7.0E−03	5.5E−04	4.9E−02	4.9E−02
SCGB1D2	11q12.3	3.5E−04	5.5E−03	2.0E−03	4.9E−02	4.9E−02
Transcriptome reference panel: TCGA breast (tumor)|primary GWAS: ER-negative BC risk (second GWAS: HGSOC risk)						
ETAA1	2p14	3.0E−03	1.5E−03	2.0E−02	2.0E−02	2.0E−02
ATP8B4	15q21.2	1.6E−03	2.2E−03	1.5E−02	2.4E−02	2.4E−02

Abbreviations: BC, breast cancer; OC, ovarian cancer; FDR, false discovery rate; ER, estrogen receptor; HGSOC, high-grade serous ovarian cancer.

consistent with this potential tumor suppressor function we observed that lower *ZNF276* expression was associated with increased breast cancer risk.

Other candidate breast cancer susceptibility genes we identified include *ESRP2* (MIM: 612960), which encodes an epithelial cell-specific regulator of splicing of the breast cancer susceptibility gene *FGFR2* (MIM: 176943)[34,35] and *SCGB1D2* (MIM: 615061), which encodes lipophilin B, which is known to be expressed in both breast and ovarian tumors.[36] Lipophilin B is tightly co-expressed with and forms a covalent complex with Mammaglobin A encoded by *SCGB2A2*, the gene next to *SCGB1D2*.[36] Mammaglobin A may be used to detect disseminated or circulating tumor cells and is under investigation as a potential immunotherapeutic target in breast cancer.[37] However, we were unable to develop gene expression prediction models for *SCGB2A2* in breast normal or tumor tissues. *BHLHA15* (MIM: 608606) encodes an estrogen-regulated transcription factor that is required to maintain mammary gland differentiation in mice,[38] and we found that decreased *BHLHA15* expression was associated with greater susceptibility to breast cancer. *ETAA1* (MIM: 613196) harbors lead variants associated at genome-wide significance with pancreatic cancer[39] and the hormone-related traits of age at menopause[40] and male-pattern baldness.[41] It encodes an activator of ATR kinase that accumulates at DNA damage sites and promotes replication fork progression and integrity.[42] Breast cancer is closely linked to DNA damage repair defects, and, in the presence of DNA damage, further loss of *ETAA1* has been shown to be synthetically lethal for the cell, suggesting that *ETAA1* expression may be essential for tumorigenesis on a background of DNA damage.[43] In keeping with this observation, we noted that elevated *ETAA1* expression was associated with increased breast cancer risk. While our pleiotropy-guided transcriptome imputation study was ongoing, a genome-wide association meta-analysis for breast cancer risk that was performed in parallel identified lead variants rs79518236 (184 kb from *BHLHA15*) and rs9712235 (244 kb from *ETAA1*) at genome-wide significance only on addition of 10,407 breast cancer cases and 7,815 controls to the Michailidou et al.[44] dataset used here. There were no known GWAS associations for breast cancer risk in these regions until the larger GWAS meta-analysis, and our concomitant identification of the same regions using gene expression imputation into a smaller GWAS underscores the power of leveraging expression data to bolster genetic discovery.

We identified 11 candidate ovarian cancer susceptibility genes (Table 2). As with breast cancer, there is strong support for a role of several genes in ovarian cancer pathogenesis, and many of these genes are in regions of the genome that harbor pleiotropic associations with other cancer types. Variants immediately upstream of *CCNE1* (MIM: 123837) are associated at genome-wide significance with bladder cancer risk.[45] *CCNE1* amplification is believed to be an early event in the development of ovarian cancer[46] and is a frequent somatic event in HGSOCs that do not carry homologous recombination DNA repair pathway defects.[47] *CCNE1* amplification is also associated with poor prognosis in triple-negative breast tumors,[48] and it is

Table 2. Candidate ovarian cancer susceptibility genes identified by pleiotropy-guided S-PrediXcan analysis.

Gene	Genomic region	p value OC	p value BC	Conditional FDR OC|BC	Conditional FDR BC|OC	Conjunction FDR
Transcriptome reference panel: GTEx ovary (normal)|primary GWAS: all invasive OC risk (second GWAS: overall BC risk)						
STRCP1	15q15.3	7.2E−04	6.4E−05	3.1E−03	8.5E−05	3.1E−03
CPNE1	20q11.22	1.2E−03	7.2E−05	5.0E−03	9.9E−05	5.0E−03
AC011330.5	15q15.3	1.7E−03	2.6E−05	5.8E−03	4.5E−05	5.8E−03
CCNE1	19q12	1.9E−03	3.2E−03	1.4E−02	4.4E−03	1.4E−02
CATSPER2P1	15q15.3	4.8E−03	1.9E−04	1.8E−02	4.1E−04	1.8E−02
UQCC1	20q11.22	3.8E−03	2.5E−03	2.8E−02	4.7E−03	2.8E−02
Transcriptome reference panel: TCGA ovary (tumor)|primary GWAS: all invasive OC risk (second GWAS: overall BC risk)						
CPNE1	20q11.22	2.0E−03	9.0E−05	2.0E−02	4.9E−04	2.0E−02
Transcriptome reference panel: GTEx ovary (normal)|primary GWAS: HGSOC risk (second GWAS: ER-negative BC risk)						
CCNE1	19q12	1.7E−03	2.0E−04	5.9E−03	1.5E−03	5.9E−03
STRCP1	15q15.3	9.2E−03	3.2E−04	3.1E−02	3.9E−03	3.1E−02
HEATR3	16q12.1	4.3E−03	3.1E−02	4.6E−02	4.4E−02	4.6E−02
Transcriptome reference panel: TCGA ovary (tumor)|primary GWAS: HGSOC risk (second GWAS: ER-negative BC risk)						
THSD7A	7p21.3	1.5E−03	1.2E−02	2.8E−02	4.3E−02	4.3E−02

Abbreviations: BC, breast cancer; OC, ovarian cancer; FDR, false discovery rate; ER, estrogen receptor.

worth noting that we observed the stronger conjunction FDR association signal for *CCNE1* in the pleiotropy-informed analysis that was based on the HGSOC and ER-negative breast cancer susceptibility GWAS datasets (Table 2). However, we noted that increased *CCNE1* expression was associated with decreased HGSOC (and ER-negative breast cancer) risk. This paradoxical direction of risk effect may be explained by the fact that *CCNE1* amplification is less common and the loss of homologous recombination (HR) pathway function is far more common in ovarian cancer, and, in the absence of a functional HR pathway, *CCNE1* is known to be essential for the developing tumor cell to survive.[49] This study suggests a role for *CCNE1* in conferring ovarian cancer risk. Intronic variants in *HEATR3* (MIM: 614951) are associated at genome-wide significance with glioma in European ancestry individuals[50] and with squamous cell esophageal carcinoma in East Asian ancestry individuals.[51] *HEATR3* was also identified by a TWAS of glioma susceptibility.[52] Intronic variants in *THSD7A* (MIM: 612249) are associated with epithelial ovarian cancer risk in East Asians,[53] albeit not at genome-wide significance (lead variant rs10260419 $p = 1 \times 10^{-7}$). Gene expression prediction models derived from breast and ovarian tissues both implicated the 15q15.3 region as a breast and ovarian cancer susceptibility region on imputation with these models into the breast and ovarian cancer GWAS data. Our analysis suggested several genes in this region (Tables 1 and 2), with the pseudogene *STRCP1* as the only common gene across breast and ovarian tissues. *STRCP1* overlaps the protein coding gene *STRC* (MIM: 606440), also identified in the breast-tissue-based analysis (Table 1), and variants in *STRC* have previously been associated with lung cancer risk (lung cancer lead variant rs35028925 $p = 2 \times 10^{-6}$).[54]

In this analysis, we chose to label the identified genes as candidate breast cancer susceptibility genes if they were identified on integrating the GTEx or TCGA breast expression prediction models with the breast cancer GWASs and incorporating pleiotropic information from the ovarian cancer GWASs and vice versa for candidate ovarian cancer susceptibility genes. However, application of the conjunction FDR over and above the conditional FDR in principle identified genes associated with both cancer types by tapping into GWAS data from both cancers. Therefore, in a sense, all these genes may well be regarded as candidate breast *and* ovarian cancer susceptibility genes. Moreover, in our pleiotropy-guided study design, the ovarian cancer dataset, in a sense, served as a replication dataset for the breast cancer findings and vice versa, which was particularly important given the lack of adequately powered and truly independent breast and ovarian cancer datasets outside of the datasets used in this study.[55]

We identified 38 candidate breast cancer susceptibility genes and 17 candidate ovarian cancer susceptibility genes in regions previously implicated by GWASs for breast cancer and ovarian cancer, respectively (Tables S3 and S5). The identification of a large number of genes in these regions is unsurprising, given that GWAS associations are the key determinant of the S-PrediXcan signal. However, we were able to take advantage of fine-scale mapping data generated by the Breast and Ovarian Cancer Association Consortia to separately pinpoint those genes where a fine-mapped candidate causal GWAS risk variant was incorporated in the PrediXcan model, suggesting that it drives the gene-based

association. Overall, we found this to be the case for 60% of the candidate susceptibility genes identified by PrediXcan in the breast and ovarian cancer susceptibility regions identified by GWASs. Comprehensive functional follow-up of the 19p13.11 breast and ovarian cancer GWAS region suggests that *ABHD8* and *ANKLE1* are the most likely targets in this region.[5] While there was no overlap between S-PrediXcan model variants for *ABHD8* and *ANKLE1* and fine-mapped risk variants in this region, S-PrediXcan did detect both genes as candidate causal susceptibility genes, with *ANKLE1* being the only gene that made the cut in both breast and ovarian tissues, suggesting that S-PrediXcan applied to pleiotropic gene-dense regions such as 19p13.11 does help short-list the key targets even in the absence of overlap with fine-mapped variants. A total of 21/38 breast and 13/17 ovarian cancer candidate susceptibility genes in the published GWAS regions were clustered at 17q21.31, reflecting the unique long-distance linkage disequilibrium structure of this region.[56] This phenomenon has also led to clustering of associations at 17q21.31 in previous TWASs of breast or ovarian cancer risk.[9,11]

Gene expression prediction models in this study were built using genomic data from women with genetically inferred European ancestry. The predictive performance of these models in a non-European ancestry cohort was not evaluated. Thus, a key limitation of this study is the potential lack of generalizability of these models to non-European ancestry cohorts. Recent analyses suggest that default TWAS models trained in large datasets such as GTEx suffer from a significant reduction in prediction accuracy, particularly in individuals of African ancestry, when compared to those of European ancestry.[57] There is an urgent and compelling need for trans-ancestry datasets that drive TWAS in diverse ancestral cohorts.

In conclusion, the powerful combination of pleiotropic breast and ovarian cancer GWAS datasets with transcriptome imputation from normal and tumor breast and ovarian tissues identified a total of 16 genomic loci (22 genes) associated with breast and ovarian cancer risks. Fine-mapping in larger GWAS datasets and deeper laboratory-based functional follow-up studies of these loci and candidate genes have the potential to provide fresh insights into the common biological underpinnings of breast and ovarian cancer.

Data and code availability

All datasets analyzed in this study are publicly available: Genome-wide summary genetic association statistics from BCAC are available at: http://bcac.ccge.medschl.cam.ac.uk/bcacdata/oncoarray/oncoarray-and-combined-summary-result/gwas-summary-results-breast-cancer-risk-2017/. Genome-wide summary genetic association statistics from OCAC are available at: https://www.ebi.ac.uk/gwas/downloads/summary-statistics (please search the GWAS catalog at the link above using the study accession numbers GCST004415 for the overall ovarian cancer and GCST004480 for the HGSOC datasets). PrediXcan prediction models trained on the GTEx version 7 data (including breast and ovarian tissues) are available here: https://zenodo.org/record/3572799. PrediXcan prediction models trained on the TCGA data (breast and ovarian tumors) are available here: https://zenodo.org/record/3818295. Code, including a tutorial, for running S-PrediXcan is available here: https://github.com/hakyimlab/MetaXcan. The data used for the analyses described in this manuscript can be obtained from dbGaP via accession number phs000424.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2021.100042.

Acknowledgments

The analyses presented in this manuscript were funded by grant number R01CA211574 from the United States National Institutes of Health/National Cancer Institute. The GTEx Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The results published here are in part based upon data generated by TCGA Research Network. The BCAC breast cancer genome-wide association analyses were supported by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research, the Ministère de l'Économie, de la Science et de l'Innovation du Québec through Génome Québec and grant PSR-SIIRI-701, the National Institutes of Health (U19 CA148065 and X01HG007492), Cancer Research UK (C1287/A10118, C1287/A16563, and C1287/A10710) and the European Union (HEALTH-F2-2009-223175, H2020 633784, and 634935). All studies and funders are listed in Michailidou et al.[22] The OCAC ovarian cancer genome-wide association meta-analyses were supported by the US National Institutes of Health (CA1X01HG007491-01 [C.I.A.], U19-CA148112 [T.A.S.], R01-CA149429 [C.M.P.], and R01-CA058598 [M.T.G.]); Canadian Institutes of Health Research (MOP-86727 [L.E.K.]), and the Ovarian Cancer Research Fund (A.B.). The COGS project was funded through a European Commission's Seventh Framework Programme grant (agreement number 223175: HEALTH-F2–2009-223175). All studies and funders are listed in Phelan et al.[23]

Declaration of interests

The authors declare no competing interests.

Received: January 3, 2021
Accepted: June 4, 2021

Web resources

BCAC, http://bcac.ccge.medschl.cam.ac.uk/bcacdata/oncoarray/oncoarray-and-combined-summary-result/gwas-summary-results-breast-cancer-risk-2017/
GTEx v.7 prediction models, https://zenodo.org/record/3572799
GWAScFDR, http://github.com/KehaoWu/GWAScFDR
MetaXcan, https://github.com/hakyimlab/MetaXcan.
NHGRI-ERB GWAS Catalog, https://www.ebi.ac.uk/gwas/downloads/summary-statistics
PredictDB Data Repository, http://predictdb.org/

TCGA, https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

TCGA breast and ovarian cancer PrediXcan models, https://zenodo.org/record/3818295

References

1. Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P.A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L.M., Ding, W., et al. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science *266*, 66–71.
2. Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., Collins, N., Gregory, S., Gumbs, C., and Micklem, G. (1995). Identification of the breast cancer susceptibility gene BRCA2. Nature *378*, 789–792.
3. Lord, C.J., and Ashworth, A. (2017). PARP inhibitors: Synthetic lethality in the clinic. Science *355*, 1152–1158.
4. Jiang, X., Finucane, H.K., Schumacher, F.R., Schmit, S.L., Tyrer, J.P., Han, Y., Michailidou, K., Lesseur, C., Kuchenbaecker, K.B., Dennis, J., et al. (2019). Shared heritability and functional enrichment across six solid cancers. Nat. Commun. *10*, 431.
5. Lawrenson, K., Kar, S., McCue, K., Kuchenbaeker, K., Michailidou, K., Tyrer, J., Beesley, J., Ramus, S.J., Li, Q., Delgado, M.K., et al.; GEMO Study Collaborators; EMBRACE; Hereditary Breast and Ovarian Cancer Research Group Netherlands (HEBON); KConFab Investigators; and Australian Ovarian Cancer Study Group (2016). Functional mechanisms underlying pleiotropic risk alleles at the 19p13.1 breast-ovarian cancer susceptibility locus. Nat. Commun. *7*, 12675.
6. Kar, S.P., Beesley, J., Amin Al Olama, A., Michailidou, K., Tyrer, J., Kote-Jarai, Z., Lawrenson, K., Lindstrom, S., Ramus, S.J., Thompson, D.J., et al.; ABCTB Investigators; AOCS Study Group & Australian Cancer Study (Ovarian Cancer); APCB BioResource; kConFab Investigators; NBCS Investigators; GENICA Network; and PRACTICAL consortium (2016). Genome-wide meta-analyses of breast, ovarian, and prostate cancer association studies identify multiple new susceptibility loci shared by at least two cancer types. Cancer Discov. *6*, 1052–1067.
7. Freedman, M.L., Monteiro, A.N.A., Gayther, S.A., Coetzee, G.A., Risch, A., Plass, C., Casey, G., De Biasi, M., Carlson, C., Duggan, D., et al. (2011). Principles for the post-GWAS functional characterization of cancer risk loci. Nat. Genet. *43*, 513–518.
8. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., Nicolae, D.L., Cox, N.J., Im, H.K.; and GTEx Consortium (2015). A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. *47*, 1091–1098.
9. Wu, L., Shi, W., Long, J., Guo, X., Michailidou, K., Beesley, J., Bolla, M.K., Shu, X.-O., Lu, Y., Cai, Q., et al.; NBCS Collaborators; and kConFab/AOCS Investigators (2018). A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. Nat. Genet. *50*, 968–978.
10. Ferreira, M.A., Gamazon, E.R., Al-Ejeh, F., Aittomäki, K., Andrulis, I.L., Anton-Culver, H., Arason, A., Arndt, V., Aronson, K.J., Arun, B.K., et al.; EMBRACE Collaborators; GC-HBOC Study Collaborators; GEMO Study Collaborators; ABCTB Investigators; HEBON Investigators; and BCFR Investigators (2019). Genome-wide association and transcriptome studies identify target genes and risk loci for breast cancer. Nat. Commun. *10*, 1741.
11. Gusev, A., Lawrenson, K., Lin, X., Lyra, P.C., Jr., Kar, S., Vavra, K.C., Segato, F., Fonseca, M.A.S., Lee, J.M., Pejovic, T., et al.; Ovarian Cancer Association Consortium (2019). A transcriptome-wide association study of high-grade serous epithelial ovarian cancer identifies new susceptibility genes and splice variants. Nat. Genet. *51*, 815–823.
12. Lu, Y., Beeghly-Fadiel, A., Wu, L., Guo, X., Li, B., Schildkraut, J.M., Im, H.K., Chen, Y.A., Permuth, J.B., Reid, B.M., et al. (2018). A Transcriptome-Wide Association Study Among 97,898 Women to Identify Candidate Susceptibility Genes for Epithelial Ovarian Cancer Risk. Cancer Res. *78*, 5419–5430.
13. Smeland, O.B., Frei, O., Shadrin, A., O'Connell, K., Fan, C.-C., Bahrami, S., Holland, D., Djurovic, S., Thompson, W.K., Dale, A.M., et al. (2019). Discovery of shared genomic loci using the conditional false discovery rate approach. Hum. Genet. *139*, 85–94.
14. Sun, L., Craiu, R.V., Paterson, A.D., and Bull, S.B. (2006). Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. Genet. Epidemiol. *30*, 519–530.
15. Yokoyama, J.S., Wang, Y., Schork, A.J., Thompson, W.K., Karch, C.M., Cruchaga, C., McEvoy, L.K., Witoelar, A., Chen, C.-H., Holland, D., et al.; Alzheimer's Disease Neuroimaging Initiative (2016). Association Between Genetic Traits for Immune-Mediated Diseases and Alzheimer Disease. JAMA Neurol. *73*, 691–697.
16. Witoelar, A., Jansen, I.E., Wang, Y., Desikan, R.S., Gibbs, J.R., Blauwendraat, C., Thompson, W.K., Hernandez, D.G., Djurovic, S., Schork, A.J., et al.; International Parkinson's Disease Genomics Consortium (IPDGC), North American Brain Expression Consortium (NABEC), and United Kingdom Brain Expression Consortium (UKBEC) Investigators (2017). Genome-wide Pleiotropy Between Parkinson Disease and Autoimmune Diseases. JAMA Neurol. *74*, 780–792.
17. Andreassen, O.A., Djurovic, S., Thompson, W.K., Schork, A.J., Kendler, K.S., O'Donovan, M.C., Rujescu, D., Werge, T., van de Bunt, M., Morris, A.P., et al.; International Consortium for Blood Pressure GWAS; Diabetes Genetics Replication and Meta-analysis Consortium; and Psychiatric Genomics Consortium Schizophrenia Working Group (2013). Improved detection of common variants associated with schizophrenia by leveraging pleiotropy with cardiovascular-disease risk factors. Am. J. Hum. Genet. *92*, 197–209.
18. Smeland, O.B., Frei, O., Kauppi, K., Hill, W.D., Li, W., Wang, Y., Krull, F., Bettella, F., Eriksen, J.A., Witoelar, A., et al.; Neuro-CHARGE (Cohorts for Heart and Aging Research in Genomic Epidemiology) Cognitive Working Group (2017). Identification of Genetic Loci Jointly Influencing Schizophrenia Risk and the Cognitive Traits of Verbal-Numerical Reasoning, Reaction Time, and General Cognitive Function. JAMA Psychiatry *74*, 1065–1075.
19. GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.
20. Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61–70.
21. Cancer Genome Atlas Research Network (2011). Integrated genomic analyses of ovarian carcinoma. Nature *474*, 609–615.
22. Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemaçon, A., Soucy, P., Glubb, D., Rostamianfar, A., et al.; NBCS Collaborators; ABCTB Investigators; and ConFab/AOCS Investigators (2017). Association analysis identifies 65 new breast cancer risk loci. Nature *551*, 92–94.

23. Phelan, C.M., Kuchenbaecker, K.B., Tyrer, J.P., Kar, S.P., Lawrenson, K., Winham, S.J., Dennis, J., Pirie, A., Riggan, M.J., Chornokur, G., et al.; AOCS study group; EMBRACE Study; GEMO Study Collaborators; HEBON Study; KConFab Investigators; and OPAL study group (2017). Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. Nat. Genet. *49*, 680–691.
24. Sankararaman, S., Sridhar, S., Kimmel, G., and Halperin, E. (2008). Estimating local ancestry in admixed populations. Am. J. Hum. Genet. *82*, 290–303.
25. Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat. Protoc. *7*, 500–507.
26. Fachal, L., Aschard, H., Beesley, J., Barnes, D.R., Allen, J., Kar, S., Pooley, K.A., Dennis, J., Michailidou, K., Turman, C., et al.; GEMO Study Collaborators; EMBRACE Collaborators; KConFab Investigators; HEBON Investigators; and ABCTB Investigators (2020). Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes. Nat. Genet. *52*, 56–73.
27. Udler, M.S., Tyrer, J., and Easton, D.F. (2010). Evaluating the power to discriminate between highly correlated SNPs in genetic association studies. Genet. Epidemiol. *34*, 463–468.
28. Lilyquist, J., Ruddy, K.J., Vachon, C.M., and Couch, F.J. (2018). Common Genetic Variation and Breast Cancer Risk-Past, Present, and Future. Cancer Epidemiol. Biomarkers Prev. *27*, 380–394.
29. Barbeira, A.N., Pividori, M., Zheng, J., Wheeler, H.E., Nicolae, D.L., and Im, H.K. (2019). Integrating predicted transcriptome from multiple tissues improves association detection. PLoS Genet. *15*, e1007889.
30. Visconti, A., Duffy, D.L., Liu, F., Zhu, G., Wu, W., Chen, Y., Hysi, P.G., Zeng, C., Sanna, M., Iles, M.M., et al. (2018). Genome-wide association study in 176,678 Europeans reveals genetic loci for tanning response to sun exposure. Nat. Commun. *9*, 1684.
31. Wong, J.C.Y., Gokgoz, N., Alon, N., Andrulis, I.L., and Buchwald, M. (2003). Cloning and mutation analysis of ZFP276 as a candidate tumor suppressor in breast cancer. J. Hum. Genet. *48*, 668–671.
32. D'Andrea, A.D. (2010). Susceptibility pathways in Fanconi's anemia and breast cancer. N. Engl. J. Med. *362*, 1909–1919.
33. Figlioli, G., Bogliolo, M., Catucci, I., Caleca, L., Lasheras, S.V., Pujol, R., Kiiski, J.I., Muranen, T.A., Barnes, D.R., Dennis, J., et al.; ABCTB Investigators; GEMO Study Collaborators; and KConFab (2019). The *FANCM*:p.Arg658* truncating variant is associated with risk of triple-negative breast cancer. NPJ Breast Cancer *5*, 38.
34. Warzecha, C.C., Sato, T.K., Nabet, B., Hogenesch, J.B., and Carstens, R.P. (2009). ESRP1 and ESRP2 are epithelial cell-type-specific regulators of FGFR2 splicing. Mol. Cell *33*, 591–601.
35. Meyer, K.B., Maia, A.-T., O'Reilly, M., Teschendorff, A.E., Chin, S.-F., Caldas, C., and Ponder, B.A.J. (2008). Allele-specific up-regulation of FGFR2 increases susceptibility to breast cancer. PLoS Biol. *6*, e108.
36. Zafrakas, M., Petschke, B., Donner, A., Fritzsche, F., Kristiansen, G., Knüchel, R., and Dahl, E. (2006). Expression analysis of mammaglobin A (SCGB2A2) and lipophilin B (SCGB1D2) in more than 300 human tumors and matching normal tissues reveals their co-expression in gynecologic malignancies. BMC Cancer *6*, 88.
37. Ghersevich, S., and Ceballos, M.P. (2014). Mammaglobin A: review and clinical utility. Adv. Clin. Chem. *64*, 241–268.
38. Zhao, Y., Johansson, C., Tran, T., Bettencourt, R., Itahana, Y., Desprez, P.-Y., and Konieczny, S.F. (2006). Identification of a basic helix-loop-helix transcription factor expressed in mammary gland alveolar cells and required for maintenance of the differentiated state. Mol. Endocrinol. *20*, 2187–2198.
39. Childs, E.J., Mocci, E., Campa, D., Bracci, P.M., Gallinger, S., Goggins, M., Li, D., Neale, R.E., Olson, S.H., Scelo, G., et al. (2015). Common variation at 2p13.3, 3q29, 7p13 and 17q25.1 associated with susceptibility to pancreatic cancer. Nat. Genet. *47*, 911–916.
40. Kichaev, G., Bhatia, G., Loh, P.-R., Gazal, S., Burch, K., Freund, M.K., Schoech, A., Pasaniuc, B., and Price, A.L. (2019). Leveraging Polygenic Functional Enrichment to Improve GWAS Power. Am. J. Hum. Genet. *104*, 65–75.
41. Yap, C.X., Sidorenko, J., Wu, Y., Kemper, K.E., Yang, J., Wray, N.R., Robinson, M.R., and Visscher, P.M. (2018). Dissection of genetic variation and evidence for pleiotropy in male pattern baldness. Nat. Commun. *9*, 5407.
42. Bass, T.E., Luzwick, J.W., Kavanaugh, G., Carroll, C., Dungrawala, H., Glick, G.G., Feldkamp, M.D., Putney, R., Chazin, W.J., and Cortez, D. (2016). ETAA1 acts at stalled replication forks to maintain genome integrity. Nat. Cell Biol. *18*, 1185–1195.
43. Achuthankutty, D., Thakur, R.S., Haahr, P., Hoffmann, S., Drainas, A.P., Bizard, A.H., Weischenfeldt, J., Hickson, I.D., and Mailand, N. (2019). Regulation of ETAA1-mediated ATR activation couples DNA replication fidelity and genome stability. J. Cell Biol. *218*, 3943–3953.
44. Zhang, H., Ahearn, T., Lecarpentier, J., Barnes, D., Beesley, J., Qi, G., Jiang, X., O'Mara, T.A., Zhao, N., Bolla, M.K., et al. (2019). Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. Nat. Genet *52*, 572–581.
45. Rothman, N., Garcia-Closas, M., Chatterjee, N., Malats, N., Wu, X., Figueroa, J.D., Real, F.X., Van Den Berg, D., Matullo, G., Baris, D., et al. (2010). A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. Nat. Genet. *42*, 978–984.
46. Karst, A.M., Jones, P.M., Vena, N., Ligon, A.H., Liu, J.F., Hirsch, M.S., Etemadmoghadam, D., Bowtell, D.D.L., and Drapkin, R. (2014). Cyclin E1 deregulation occurs early in secretory cell transformation to promote formation of fallopian tube-derived high-grade serous ovarian cancers. Cancer Res. *74*, 1141–1152.
47. Bowtell, D.D.L. (2010). The genesis and evolution of high-grade serous ovarian cancer. Nat. Rev. Cancer *10*, 803–808.
48. Zhao, Z.-M., Yost, S.E., Hutchinson, K.E., Li, S.M., Yuan, Y.-C., Noorbakhsh, J., Liu, Z., Warden, C., Johnson, R.M., Wu, X., et al. (2019). CCNE1 amplification is associated with poor prognosis in patients with triple negative breast cancer. BMC Cancer *19*, 96.
49. Etemadmoghadam, D., Weir, B.A., Au-Yeung, G., Alsop, K., Mitchell, G., George, J., Davis, S., D'Andrea, A.D., Simpson, K., Hahn, W.C., Bowtell, D.D.; and Australian Ovarian Cancer Study Group (2013). Synthetic lethality between CCNE1 amplification and loss of BRCA1. Proc. Natl. Acad. Sci. USA *110*, 19489–19494.
50. Melin, B.S., Barnholtz-Sloan, J.S., Wrensch, M.R., Johansen, C., Il'yasova, D., Kinnersley, B., Ostrom, Q.T., Labreche, K., Chen, Y., Armstrong, G., et al.; GliomaScan Consortium (2017).

Genome-wide association study of glioma subtypes identifies specific differences in genetic susceptibility to glioblastoma and non-glioblastoma tumors. Nat. Genet. *49*, 789–794.

51. Wu, C., Kraft, P., Zhai, K., Chang, J., Wang, Z., Li, Y., Hu, Z., He, Z., Jia, W., Abnet, C.C., et al. (2012). Genome-wide association analyses of esophageal squamous cell carcinoma in Chinese identify multiple susceptibility loci and gene-environment interactions. Nat. Genet. *44*, 1090–1097.
52. Atkins, I., Kinnersley, B., Ostrom, Q.T., Labreche, K., Il'yasova, D., Armstrong, G.N., Eckel-Passow, J.E., Schoemaker, M.J., Nöthen, M.M., Barnholtz-Sloan, J.S., et al. (2019). Transcriptome-Wide Association Study Identifies New Candidate Susceptibility Genes for Glioma. Cancer Res. *79*, 2065–2071.
53. Lawrenson, K., Song, F., Hazelett, D.J., Kar, S.P., Tyrer, J., Phelan, C.M., Corona, R.I., Rodríguez-Malavé, N.I., Seo, J.-H., Adler, E., et al.; Australian Ovarian Cancer Study Group (2019). Genome-wide association studies identify susceptibility loci for epithelial ovarian cancer in east Asian women. Gynecol. Oncol. *153*, 343–355.
54. McKay, J.D., Hung, R.J., Han, Y., Zong, X., Carreras-Torres, R., Christiani, D.C., Caporaso, N.E., Johansson, M., Xiao, X., Li, Y., et al.; SpiroMeta Consortium (2017). Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nat. Genet. *49*, 1126–1132.
55. Huffman, J.E. (2018). Examining the current standards for genetic discovery and replication in the era of mega-biobanks. Nat. Commun. *9*, 5054.
56. de Jong, S., Chepelev, I., Janson, E., Strengman, E., van den Berg, L.H., Veldink, J.H., and Ophoff, R.A. (2012). Common inversion polymorphism at 17q21.31 affects expression of multiple genes in tissue-specific manner. BMC Genomics *13*, 458.
57. Keys, K.L., Mak, A.C.Y., White, M.J., Eckalbar, W.L., Dahl, A.W., Mefford, J., Mikhaylova, A.V., Contreras, M.G., Elhawary, J.R., Eng, C., et al. (2020). On the cross-population generalizability of gene expression prediction models. PLoS Genet. *16*, e1008927.

HGGA, Volume *2*

Supplemental information

Pleiotropy-guided transcriptome imputation from normal and tumor tissues identifies candidate susceptibility genes for breast and ovarian cancer

Siddhartha P. Kar, Daniel P.C. Considine, Jonathan P. Tyrer, Jasmine T. Plummer, Stephanie Chen, Felipe S. Dezem, Alvaro N. Barbeira, Padma S. Rajagopal, Will T. Rosenow, Fernando Moreno, Clara Bodelon, Jenny Chang-Claude, Georgia Chenevix-Trench, Anna deFazio, Thilo Dörk, Arif B. Ekici, Ailith Ewing, George Fountzilas, Ellen L. Goode, Mikael Hartman, Florian Heitz, Peter Hillemanns, Estrid Høgdall, Claus K. Høgdall, Tomasz Huzarski, Allan Jensen, Beth Y. Karlan, Elza Khusnutdinova, Lambertus A. Kiemeney, Susanne K. Kjaer, Rüdiger Klapdor, Martin Köbel, Jingmei Li, Clemens Liebrich, Taymaa May, Håkan Olsson, Jennifer B. Permuth, Paolo Peterlongo, Paolo Radice, Susan J. Ramus, Marjorie J. Riggan, Harvey A. Risch, Emmanouil Saloustros, Jacques Simard, Lukasz M. Szafron, Linda Titus, Cheryl L. Thompson, Robert A. Vierkant, Stacey J. Winham, Wei Zheng, Jennifer A. Doherty, Andrew Berchuck, Kate Lawrenson, Hae Kyung Im, Ani W. Manichaikul, Paul D.P. Pharoah, Simon A. Gayther, and Joellen M. Schildkraut

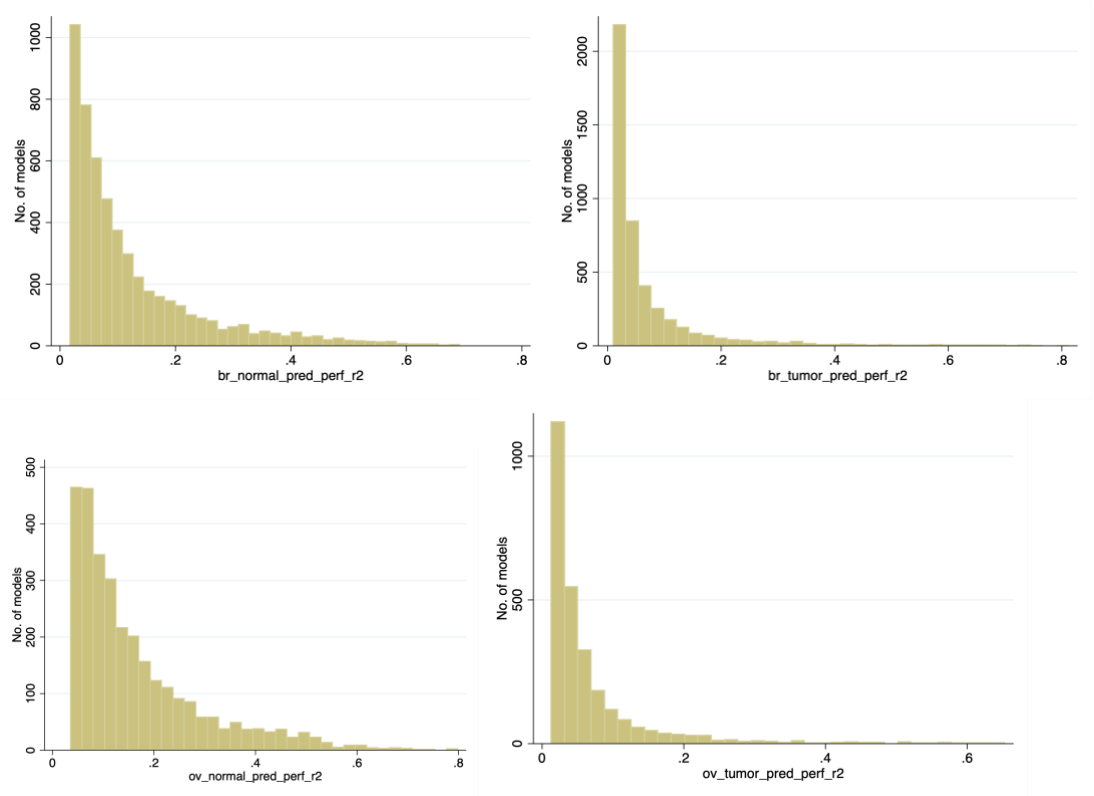

Figure S1: Prediction performance of PrediXcan models in normal and tumor breast and ovarian tissues.

Pred_perf_r2 is the cross-validated R2 of the tissue model's correlation to the gene's measured transcriptome (prediction performance). This prediction performance is displayed for the GTEx normal breast and ovarian tissues and the TCGA breast and ovarian tumor tissues, demonstrating that, in general, prediction performance was better for the GTEx normal tissues than the TCGA tumor tissues.

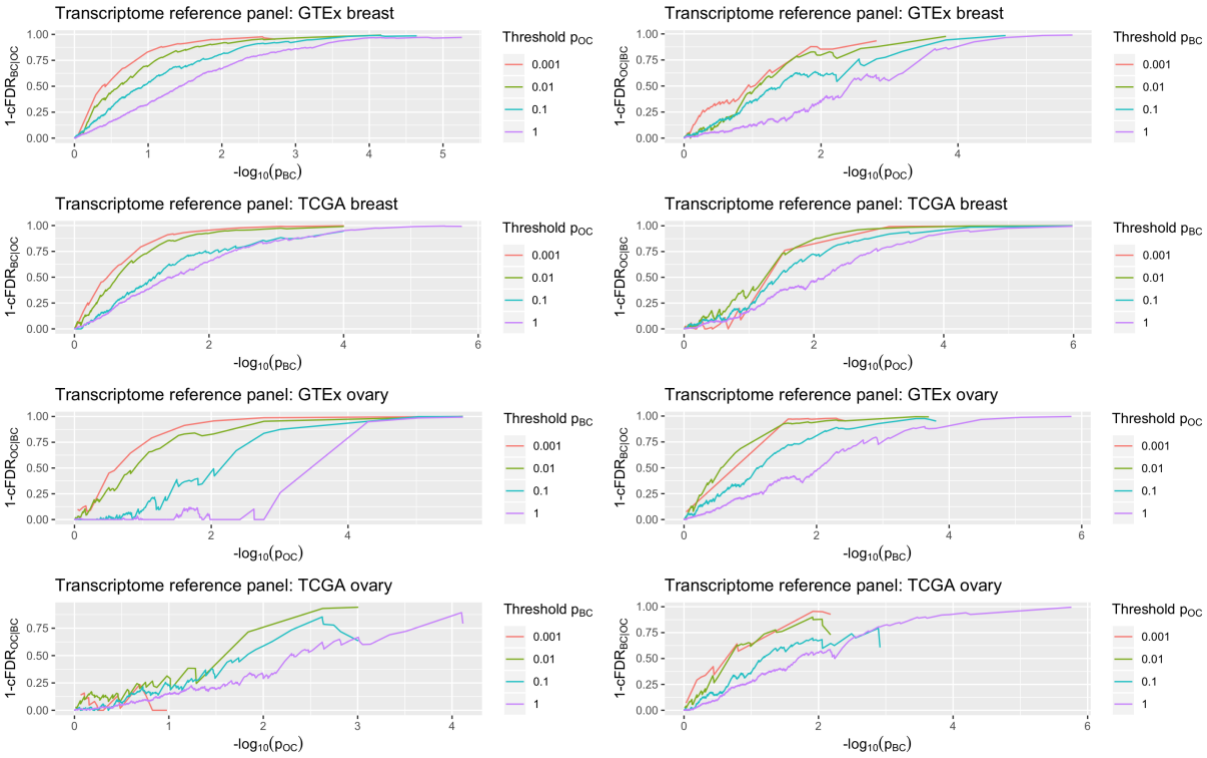

Figure S2: True discovery rate of S-PrediXcan associations for each cancer stratified by associations with the other cancer for the subtype-specific analyses.

True discovery rate against the negative logarithm (base 10) of the *P*-value for each cancer for subsets of genes based on strength of association with the other cancer for the subtype-specific analyses. The Y-axis of each plot is the true discovery rate which is defined as 1 – conditional false discovery rate (cFDR). For a given ordered analytic combination of data sets (e.g., GTEx normal breast tissue as transcriptome reference panel-estrogen receptor (ER)-negative breast cancer GWAS-high-grade serous ovarian cancer (HGSOC) GWAS, plotted in the upper left hand corner) we observed that, in general, for progressively smaller S-PrediXcan *P*-values of the second cancer type (indicated by the key "Threshold p" next to each plot), the true discovery rate (Y-axis) for association with the primary cancer type approached 100% at progressively larger S-PrediXcan *P*-values for the primary cancer type (X-axis; negative logarithm (base 10) of the *P*-values). BC: ER-negative breast cancer risk; OC: HGSOC risk. Only *P*-values > 10^{-6} are plotted on the X-axis.